CHAPTER 5

Correlation and Regression

Summary _____

Correlation and *regression* are two different statistical methods that are closely related mathematically. Both methods provide information about *bivariate distributions*. A bivariate distribution has two variables whose scores are logically paired.

Correlation and regression are used for different purposes. A correlation coefficient (symbolized as *r*), is used to describe the *degree* and the *direction* of a relationship between two variables. A regression equation is used to draw a *line of best fit* and to *predict scores* on one variable, if you have scores on the other variable. The textbook describes the simplest case for these two methods, the case which requires that the two variables have a *linear relationship*. You can make a visual check for linearity by constructing and examining a scatterplot of the two variables. This is a recurring theme in the text and the study guide; when in doubt, graph the data to gain insight into the nature of the relationship between variables.

The two most important names in this chapter are Francis Galton, who conceived the idea of a correlation statistic, and Karl Pearson, who developed the mathematics for calculating the correlation coefficient.

Your textbook gives two methods of arranging the arithmetic when you calculate *r*. If you choose to use the "blanched" (partially cooked) formula, be sure you (or your calculator) carry three or four decimal places in the calculations.

To summarize several of the characteristics of *r* and its interpretation:

a. The *algebraic sign* of *r* gives the direction of the relationship. If the sign is positive, the relationship is direct (higher scores on one variable go with higher scores on the other variable). If the sign is negative, the relationship is inverse (higher scores on one variable go with lower scores on the other).

b. The closer the absolute value of r is to 1.00, the stronger the relationship, and the more confidence you can put into a prediction made from a regression equation based on the data.

c. Positive coefficients are not "better" than negative coefficients.

d. r^2, the *coefficient of determination*, gives the proportion of variance the two variables have in common.

e. Correlation coefficients, no matter how large, *are not sufficient evidence to claim a causal relationship between two variables.*

f. Low correlations do not necessarily mean that there is no relationship between the two variables; *nonlinear relationships* and *truncated ranges* both produce spuriously (artificially) low correlation coefficients.

g. The *effect size index* for a correlation coefficient is the correlation coefficient itself. Depending on the reason that r was calculated, descriptive adjectives of *small, medium, and large* are appropriate for different values of r.

h. When the same measure is administered twice to the same subjects, a correlation coefficient of .80 or greater indicates that the measure is *reliable.*

Besides the Pearson product-moment correlation coefficient, which is used for *two quantitative variables*, other kinds of correlation coefficients are used when the relationship between variables is examined. Any correlation coefficient expresses the strength and direction of the relationship between variables.

A *regression equation*, $\hat{Y} = a + bX$, allows you to predict a value for Y for any value of X. The prediction will be more accurate when the relationship between X and Y is linear and the correlation coefficient is large.

To write a regression equation for a bivariate distribution, calculate values for the *two regression coefficients*, a and b. The regression coefficient, a is the *intercept* of the regression line with the Y axis, and the coefficient b is the slope of the regression line.

A regression line can be presented as a graph, but its appearance will depend on he units used on the X and Y axes. Also, there are two regression lines for one set of ›ivariate data. The line that your calculations produce depends on which variable ou designate as the Y variable.

Iultiple-Choice Questions

. To use the regression equation technique described in your text, you must have
(1) a logical pairing of the scores on the two variables;
(2) a linear relationship between the two variables;
(3) both (1) and (2);
(4) neither (1) nor (2).

Quantification is the idea that
(1) all things can be counted;
(2) all physical things can be counted;
(3) the numerical representation of a phenomenon gives the most important picture;
(4) a phenomenon can be better understood if its important parts are expressed as numbers.

A Pearson correlation coefficient is appropriate to describe which of the situations ›w?
(1) As X increases, Y decreases by the same amount;
(2) As X increases, Y goes up at first slowly and then faster;
(3) As X increases, Y goes up at first and then goes down;
(4) All of the above.

correlation of -.88 between television viewing time and grades in high school is best understood as demonstrating that
(1) as television viewing time increases, grades increase;
(2) as television viewing time decreases, grades decrease;
(3) as television viewing time increases, grades decrease;
(4) both (1) and (2) are correct.

5. Assume you conduct a study to evaluate the relationship between the amount of time a child is read to and reading ability at age 15. You find a correlation coefficient of .04. This correlation suggests
   (1) a very strong relationship; as reading to children increases, reading ability increases;
   (2) a very weak relationship; as reading to children increases, reading ability decreases slightly;
   (3) almost no relationship;
   (4) unable to tell without knowing the number of participants.

6. A linear relationship is described by which of the statements below?
   (1) The two variables are paired in some logical fashion;
   (2) For every one-point increase in one variable, you get a four-point increase in the other variable;
   (3) Both (1) and (2);
   (4) Neither (1) nor (2).

7. A Pearson product-moment correlation coefficient can be used to express the degree of relationship for which situation(s) below?
   (1) A little anxiety produces poor results, a moderate amount produces good results, and a high level of anxiety produces poor results;
   (2) Early in training each trial helps only a little, but as training progresses, each trial causes larger and larger gain;
   (3) For every extra year of growth in a pine forest, you can expect an increase of 10,000 board feet;
   (4) All of the above.

8. Which of the following statements is (are) true?
   (1) Correlations range from -1 to +1;
   (2) Correlations show causal relationships;
   (3) Correlations allow us to evaluate the strength of a relationship;
   (4) Both (1) and (3) are correct.

9. Psychologists have demonstrated that number of hours spent in class is correlated with grades in that class. The correlation between number of hours in class and grades is

(1) positive;
(2) negative;
(3) zero;
(4) not determinable from the information given.

10. Identify the incorrect statement.

(1) A negative correlation is obtained when high scores on X go with low scores on Y and low scores on X go with high scores on Y;
(2) A positive correlation is obtained when high scores on X go with high scores on Y;
(3) A zero correlation is obtained when high scores on X go with both high and low scores on Y and low scores on X go with both high and low scores on Y;
(4) None of the above.

11. The coefficient of determination allows you to

(1) determine the variance two variables have in common;
(2) draw cause-and-effect statements;
(3) predict X scores, given Y scores;
(4) quickly determine the regression coefficients.

12. Given a correlation coefficient of zero, which conclusion is correct?

(1) There is no relationship between the two variables;
(2) A correlation coefficient is not proper for the data;
(3) Correlation coefficients of zero cannot be interpreted;
(4) All of the above.

13. The least squares method of finding a formula for a straight line

(1) produces a slope and an intercept;
(2) makes the error in prediction a minimal amount;
(3) was championed by Karl Pearson;
(4) all of the above.

Chapter 5

14. Pearson product-moment correlation coefficients can be used to establish the degree of relationship
    (1) even if the two variables are measuring different things;
    (2) even if the full ranges of the two variables are not included in the data;
    (3) even though the relationship is not linear;
    (4) all of the above.

15. The regression coefficient, a, is most clearly related to
    (1) the angle the regression line makes with the X axis;
    (2) the place the regression line crosses the Y axis;
    (3) the Y score predicted for the X score that is the mean of the X distribution;
    (4) the absolute size of the correlation coefficient.

16. One reason for a small correlation might be
    (1) truncated range;
    (2) nonlinear relationship;
    (3) neither (1) nor (2);
    (4) both (1) and (2).

17. Suppose you know that the regression coefficients for the line that predicts the number of offspring of children from the number of offspring of the maternal grandmother are $a = 10$, $b = -1.0$. Knowing this, you can conclude that the correlation coefficient for these data is
    (1) positive;
    (2) negative;
    (3) perfect;
    (4) none of the above.

18. Error in a regression analysis is defined as
    (1) $\hat{Y}$;
    (2) $Y - \hat{Y}$;
    (3) $\hat{Y} - Y$;
    (4) $(Y - \hat{Y})^2$.

Chapter 5

19. Suppose you had the exam scores on the first hour exam for 100 general psychology students. A correlation coefficient could be calculated if the scores were divided according to the variable_____.
    (1) gender-------males and females;
    (2) where a person sits in the class-------front or back;
    (3) both (1) and (2);
    (4) neither (1) nor (2).

20. The difference between correlation coefficients and regression equations is that correlation coefficients
    (1) allows us to infer causation; regression equations do not;
    (2) allows prediction of scores; regression equations do not;
    (3) give the relationship between two variables; regression equations predict scores;
    (4) they are the same procedure.

Short-Answer Questions_____

1. "The self-confidence of that group of recruits is negatively correlated with their success in the obstacle course." Tell what this statement means.

2. Describe the statistical method of regression. Tell what it is good for and what its limitations are.

3. A study of 4138 students in 25 law schools found a correlation coefficient of .36 between first-year law school grades and scores on the Law School Admission Test. Interpret the meaning of this correlation.

4. Assume that the correlation between SAT scores and GPA in college is .35. Interpret the meaning of this correlation.

5. Some personality assessments ask you to identify items in a series that are "like you" or "not like you." Examples are items such as, "I am a hard worker" or "I don't care what happens to my neighbors." One of the problems facing psychologists who are developing personality assessments is that the items differ in their social desirability. (The two items above are good examples). Asking people to rate the

social desirability of each item on a scale of 1 to 9 is a common way to find out about items.

Imagine that 20 nuns rate a set of 100 items. For each item a mean is calculated Now imagine that 20 convicted child molesters rate the same set of items. Again a mean is calculated for each item. Stop a moment and estimate the correlation coefficient for the two sets of ratings (positive or negative, high or low). Based on other studies, the correlation coefficient between the two sets of ratings will be abc .90.

a. What is the N that the correlation is based on?
b. Write a sentence that explains the meaning of this correlation coefficient.

Problems

1. The idea of using tests to predict who will do well in college began to emerge around 1900. Many (including Galton) assumed that people with quick reaction times and keen sensory abilities would be quick thinkers with keen intellects (who would, of course, make good grades). James McKeen Cattell at Columbia Universit gathered data on this assumption. The summary statistics below are representative c his findings, as reported by Clark Wissler. (See Sokal [1982] for an overview.)

Sensory Ability Score	Grade Point Average
$\Sigma X = 3500$	$\Sigma Y = 115$
$\Sigma X^2 = 250{,}000$	$\Sigma Y^2 = 289$

$\Sigma XY = 8084$
$N = 50$

a. Calculate the correlation coefficient between sensory ability and freshman grade point average.
b. Write the regression equation that predicts GPA.
c. Predict the GPA of a person with a sensory ability score of 100.
d. Tell in words how much faith you have that a person with a sensory ability score of 100 would have the GPA you predicted.

Chapter 5

Does reading improve vocabulary? An elementary school teacher thought so. To :ove the case, he gathered data for 12 children on time spent reading. Each child's me score in minutes matched with his or her vocabulary score. Summary statistics)llow.

eading Time	Vocabulary Score
$\Sigma X = 252$	$\Sigma Y = 204$
$\Sigma X^2 = 7644$	$\Sigma Y^2 = 3607$
$\Sigma XY = 4787$	
$N = 12$	

a. Find the correlation coefficient between reading time and vocabulary score.
b. Write the regression equation that predicts vocabulary score.
c. Predict the vocabulary score of a child with a reading time of 50 minutes.
d. Tell in words how much faith you have that a child with a reading time of 50 minutes would have the vocabulary score you predicted.
e. Comment on this conclusion: "These data show that reading improves vocabulary."

. Is creativity related to humor? Each student in this data set has a score on a test of creativity and a score based on the number of puns produced while looking at a list of "wise sayings." Draw a scatterplot and find the correlation coefficient and the coefficient of determination. Write the regression equation and plot the line on your scatterplot. Write an explanation of what your analysis shows. Predict the number of puns for a student whose creativity test score was 93.

;tudent	Creativity Test Score	Number of Puns
1	60	28
2	57	32
3	52	24
4	46	16
5	41	21
6	38	14
7	32	18
8	29	11
9	25	9
10	19	12

Chapter 5

4. Studies of conformity require a participant to make a judgment about a stimulus. Afterward, pressure to conform is applied (which might be that everyone else gives a judgment that differs from the participant's judgment). The stimulus (or one similar) is presented again. The dependent variable is the amount of change in the participant's judgment.

Each subject below participated in two conformity sessions. The first session involved estimates of distance; the second involved value judgments about the degree of truth in ten controversial statements. Begin by looking at the data and making a judgment about the size and direction of the correlation coefficient. (To get a feeling for being a participant in conformity research, share your estimate with fellow class members, who then share theirs with you.) Calculate an *r* and write an interpretation. Write the regression equation for these data, treating the value judgments as the *Y* variable.

Participant	Distance Judgments	Value Judgments
1	8	1
2	4	2
3	7	0
4	9	3
5	3	1
6	0	2
7	4	0

5. Assume you are interested in predicting the GPA of college students by knowing their high school class rank. From records, you are able to obtain the high school ranks of 8 students, and they you find those students' college GPAs. From these data, create a scatter plot, write a regression equation and draw your line on the plot. Assume you have a student with a high school rank of 4. Predict the student's college GPA.

Participant	High School Rank	College GPA
1	28	3.16
2	4	2.55
3	47	3.10
4	19	3.97
5	43	1.84
6	50	2.99
7	34	2.07
8	21	3.01